Ann. Inst. Statist. Math.
29 (1977), Part A, 153-164

AN EXTENSION OF THE METHOD OF MAXIMUM LIKELIHOOD AND THE STEIN'S PROBLEM*

HIROTUGU AKAIKE

(Received Mar. 25, 1976)

Summary

An extension of the method of maximum likelihood leads to a natural solution of the problem raised by Stein, the inadmissibility of the ordinary maximum likelihood estimator for the mean of a multivariate normal distribution.

1. Introduction

By explicitly constructing a superior estimator Stein [12] showed that the ordinary maximum likelihood estimator of the mean of a multivariate normal distribution is inadmissible in terms of the sum of the mean squared errors. James and Stein [8] introduced another estimator of which improvement over the maximum likelihood estimator could be quite substantial. These estimators are obtained by multiplying the maximum likelihood estimator by properly chosen shrinkage factors and we will call them by a generic name, the Stein estimator. The further improved estimators obtained by taking the positive parts of the shrinkage factors (Stein [14]) will be called the positive part estimator.

Since the first paper by Stein [12] it has been customary to motivate the new estimators by a Bayesian reasoning (Stein [13], [14], Dempster [4], Efron and Morris [5], [6]). Lindley [10] even noted that it was difficult to see how someone who would wish us to use only the likelihood function could use anything other than the simple maximum likelihood estimator in the situation treated by Stein. Certainly Bayesian approaches can be quite useful for the design and assessment of an estimator with a given loss function but it quickly becomes obvious that they can not fully explain the nature of the Stein estimator. The explanation based on an empirical Bayesian approach (Stein [12]) depends heavily on the asymptotic property of the sequence of the parameters to be

* Part of this paper was presented at an invited session of the Fall Meeting of the Mathematical Society of Japan, Tokyo, October 1975.

estimated, while the dimension of the vector of the parameters may be as small as 3 or 4. As is explained by Lindley [10], subjective Bayesians who are not so bold as to ignore the information obtained through the observations might accept the concept of the Stein estimator, but they are, at least at present, in need of what Good [7] calls a Bayes/non-Bayes compromise in determining a parameter within the prior distribution. In this paper it is shown that by a careful examination of the method of maximum likelihood we can develop an estimation procedure without recourse to the Bayesian argument.

The main idea of this paper is to treat the log likelihood as an estimate of the expected log likelihood which is, except for an additive constant, equal to the generalized entropy of the true distribution with respect to the distribution defined by the assumed density function. The generalized entropy is equal to the minus of the Kullback information quantity which measures the discrepancy between the two distributions. The value of the log likelihood function at the maximum likelihood estimates of the parameters is always biased to overestimate the generalized entropy. Thus if we consider the aim of our estimating the parameters to be the use of the probability distribution, defined by the estimates, for the prediction of future observations, then, contrary to the common understanding of the unbiasedness of the maximum likelihood estimator of the mean of a multivariate normal distribution, the maximum likelihood estimator must be considered to be biased towards the direction to cause the apparent increase of the log likelihood which is, ignoring the additive constant, a natural estimate of the generalized entropy. In the case treated by Stein the generalized entropy of the true distribution with respect to the estimated distribution is identical to a half of the minus of the sum of the squared deviations of the estimated means from the true means. Thus the optimization with respect to the mean squared error criterion is equivalent to the optimization with respect to the expected generalized entropy. The above mentioned overestimating characteristic of the maximum likelihood estimator may be considered as a manifestation of the over sensitivity of the estimator to the variation of the sample values and the shrinking appears to be a natural countermeasure. In trying to determine the optimum shrinkage factor we find that the only quantity to be estimated from the data is the sum of the squares of the true means. The estimate is obtained by the method of maximum likelihood. Thus the whole process is closed and complete without calling for any ad hoc principles. It turns out that the new estimator thus obtained is close to some of the Stein positive part estimators.

2. Log likelihood and entropy

For a set of independently and identically distributed random variables x_i $(i=1, 2, \cdots, N)$ the average log likelihood $(1/N)\sum \log f(x_i|\boldsymbol{\theta})$ of a probability density function $f(x|\boldsymbol{\theta})$ with respect to a measure dx is a natural estimate of the expected log likelihood $\mathrm{E}\log f(x|\boldsymbol{\theta})$. If we denote the probability density function of x_i by $g(x)$, $\mathrm{E}\log f(x|\boldsymbol{\theta}) = \int g(x)\log f(x|\boldsymbol{\theta})dx$.

The generalized entropy of the distribution $g(x)$ with respect to the distribution $f(x|\boldsymbol{\theta})$ is defined by

$$-\int \left(\frac{g(x)}{f(x|\boldsymbol{\theta})}\right)\log\left(\frac{g(x)}{f(x|\boldsymbol{\theta})}\right)f(x|\boldsymbol{\theta})dx\,.$$

The probabilistic interpretation of the thermodynamical concept of entropy, developed by Boltzmann [3], as the logarithm of the probability of getting a sample distribution from a basic distribution provided the basis for the development of quantum physics at the beginning of this century and the generalized entropy retains a significant status as a most natural measure of deviation of $g(x)$ from $f(x|\boldsymbol{\theta})$ (see Sanov [11], Vincze [15]). The minus of this generalized entropy is known as Kullback's information quantity and its analytic properties are extensively discussed by Kullback [9]. Since the generalized entropy can be expressed in the form $\mathrm{E}\log f(x|\boldsymbol{\theta}) - \mathrm{E}\log g(x)$, the method of maximum likelihood may be viewed as a method of finding a $\boldsymbol{\theta}$ which will tend to maximize $\mathrm{E}\log f(x|\boldsymbol{\theta})$, or the generalized entropy, through the observations of x_i. This interpretation of the method of maximum likelihood was first developed by the present author in 1971 and lead to the introduction of an information criterion (AIC) for the comparison of statistical models fitted by the method of maximum likelihood (Akaike [1], [2]).

Now consider the situation treated by Stein. In this case the observations z_i are independently distributed as $N(\xi_i, 1)$ $(i=1, 2, \cdots, p)$. Since ξ_i's are not necessarily identical we can not apply the foregoing discussion of entropy maximizing property of the method of maximum of likelihood directly to the present situation. Here we consider $\boldsymbol{x}$ to be a p-dimensional vector $(x_1, x_2, \cdots, x_p)'$ and assume that $\boldsymbol{z}=(z_1, z_2, \cdots, z_p)'$ is the result of observing $\boldsymbol{x}$ once. We are assuming the situation where $\log g(x) = (-1/2)\{p\log 2\pi + \sum (x_i-\xi_i)^2\}$ and $\log f(x|\boldsymbol{\theta}) = (-1/2)\cdot\{p\log 2\pi + \sum (x_i-\theta_i)^2\}$. From the definition the generalized entropy of $g(x)$ with respect to $f(x|\boldsymbol{\theta})$ is given by $(-1/2)\sum (\theta_i-\xi_i)^2$. Thus the sum of the squared errors appears as a natural criterion of fit of $f(x|\boldsymbol{\theta})$ to $g(x)$.

One might wish to use $B=\log f(z|\theta)-\mathrm{E}\log g(x)$ as an estimate of the generalized entropy. We have

$$B=(p/2)\{1-(1/p)\sum(z_i-\theta_i)^2\}\ .$$

Maximizing the log likelihood is equivalent to maximizing B with respect to θ_i and the maximum $p/2$ of B is attained at $\theta_i=z_i$ $(i=1, 2, \cdots, p)$. It is well known that Kullback's information quantity is non-negative, i.e., the generalized entropy is non-positive. Thus the present estimate B is definitely an overestimate of the generalized entropy and is obviously useless as a criterion of fit, since it is completely insensitive to the deviation of θ_i from ξ_i. The expected generalized entropy of $g(x)$ with respect to $f(x|\theta)$ with $\theta=z$ is given by $(-1/2)\sum \mathrm{E}(z_i-\xi_i)^2$, where E denotes the expectation with respect to the distribution of z. This quantity is equal to $-p/2$. This shows that the expected lack of fit of a distribution defined by the ordinary maximum likelihood estimator is equal to the degree of overfit observed by the value of B.

Through the rest of the present paper we will adopt the convention to denote a vector of parameters or a random variable by a bold face letter. The values taken by a random variable $\boldsymbol{x}$ will be denoted by x.

3. Maximizing the entropy

Motivated by the observations of the preceding section we proceed to the problem of choosing a shrinkage factor ρ such that the expected generalized entropy is maximized by replacing the maximum likelihood estimator $\boldsymbol{z}$ by $\rho\boldsymbol{z}$. This is to find a ρ which minimizes the sum of the mean squared errors $\mathrm{E}\sum(\xi_i-\rho z_i)^2$. For a given $\boldsymbol{z}$ the ρ which minimizes $\|\boldsymbol{\xi}-\rho\boldsymbol{z}\|^2$ is given by $\rho_{\mathrm{opt}}=(\boldsymbol{\xi}, \boldsymbol{z})/\|\boldsymbol{z}\|^2$, where $(\boldsymbol{\xi}, \boldsymbol{z})$ denotes the inner product $\sum\xi_i z_i$ and $\|\boldsymbol{z}\|^2=(\boldsymbol{z}, \boldsymbol{z})$. This ρ_{opt} can be factored into the form

$$\rho_{\mathrm{opt}}=\frac{(\boldsymbol{\xi}, \boldsymbol{z})}{\|\boldsymbol{\xi}\|\cdot\|\boldsymbol{z}\|}\cdot\frac{\|\boldsymbol{\xi}\|}{\|\boldsymbol{z}\|}\ .$$

To get the exact value of ρ_{opt} the information of both $(\boldsymbol{\xi}, \boldsymbol{z})$ and $\|\boldsymbol{\xi}\|$ is required. Consider the situation where only $\|\boldsymbol{\xi}\|$ is known. For this case the best choice of ρ is given by

$$\rho_c=\mathrm{E}\frac{(\boldsymbol{\xi}, \boldsymbol{z})}{\|\boldsymbol{\xi}\|\cdot\|\boldsymbol{z}\|}\cdot\frac{\|\boldsymbol{\xi}\|}{\|\boldsymbol{z}\|}\ , \tag{3.1}$$

where E denotes the conditional expectation given $\|\boldsymbol{z}\|$. The joint density $g(u, w)$ of $u=(\boldsymbol{\xi}, \boldsymbol{z})/\|\boldsymbol{\xi}\|$ and $w=\|\boldsymbol{z}\|^2$ is already given by James and Stein [8] as

$$\frac{(w-u^2)^{(p-3)/2}}{(2\pi)^{1/2}2^{(p-1)/2}\Gamma((p-1)/2)}\exp\left\{-\frac{1}{2}\|\boldsymbol{\xi}\|^2+\|\boldsymbol{\xi}\|u-\frac{1}{2}w\right\}$$

if $u^2 \leqq w$ and 0 elsewhere. Thus the conditional density $g(u|w)$ of $\boldsymbol{u}$ conditional on $\boldsymbol{w}=w$ is given by

$$g(u|w)=C(w-u^2)^{(p-3)/2}\exp(\|\boldsymbol{\xi}\|u)$$

for $u^2 \leqq w$ and 0 elsewhere and C is a normalizing constant. With the change of variable $u=tw^{1/2}$ the conditional expectation of $(\boldsymbol{\xi}, \boldsymbol{z})/(\|\boldsymbol{\xi}\| \cdot \|\boldsymbol{z}\|)$ given $\|\boldsymbol{z}\|=w^{1/2}$ is obtained as

$$\frac{\int_{-1}^{1} t(1-t^2)^{(p-3)/2}\exp(\|\boldsymbol{\xi}\|w^{1/2}t)dt}{\int_{-1}^{1}(1-t^2)^{(p-3)/2}\exp(\|\boldsymbol{\xi}\|w^{1/2}t)dt}. \tag{3.2}$$

Using the Bessel function $I_\nu(s)$ which satisfies the relations (Watson [16], p. 79)

$$I_\nu(s)=\frac{((1/2)s)^\nu}{\Gamma(\nu+1/2)\Gamma(1/2)}\int_{-1}^{1}(1-t^2)^{\nu-1/2}\exp(st)dt$$

and

$$\frac{d}{ds}I_\nu(s)-\frac{\nu}{s}I_\nu(s)=I_{\nu+1}(s),$$

the above conditional expectation (3.2) can be expressed in a compact form

$$\frac{I_{\nu+1}(\|\boldsymbol{\xi}\|\cdot\|\boldsymbol{z}\|)}{I_\nu(\|\boldsymbol{\xi}\|\cdot\|\boldsymbol{z}\|)}, \tag{3.3}$$

where $\nu=(p-2)/2$. Note that we are using $\|\boldsymbol{z}\|$ for $w^{1/2}$. We denote the above quantity by ANGFTR$(\|\boldsymbol{\xi}\|\cdot\|\boldsymbol{z}\|)$. Here ANGFTR stands for the angle factor, the part of our shrinkage factor which is determined by the average angular relation between $\boldsymbol{\xi}$ and $\boldsymbol{z}$ when $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{z}\|$ are given. Incidentally, from the definition of ANGFTR, its absolute value cannot be greater than 1 and we get a proof of the fact that $I_\nu(s) \geqq I_{\nu+1}(s)$ for $s \geqq 0$. Obviously ANGFTR is non-negative.

ANGFTR$(\|\boldsymbol{\xi}\|\cdot\|\boldsymbol{z}\|)$ multiplied by the ratio $\|\boldsymbol{\xi}\|/\|\boldsymbol{z}\|$ gives the best shrinkage factor under the condition that the factor depends only on $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{z}\|$. Now the last and the most serious problem is how to get an estimate of $\|\boldsymbol{\xi}\|$ which will give a good shrinkage factor when it is used here in place of $\|\boldsymbol{\xi}\|$. This corresponds to the stage where the Bayes/non-Bayes compromise is required in the Bayesian approach and usually some ad hoc criterion such as unbiasedness is invoked for the choice of the estimate. Contrary to this conventional approach we consistently apply the method of maximum likelihood for the estimation of $\|\boldsymbol{\xi}\|$. This stage of estimation of $\|\boldsymbol{\xi}\|$ necessitates, and is made possible

only by, the aggregate use of the observations z_i. The distribution of $w=\|z\|^2$ is a noncentral chi-squared with p degrees of freedom and noncentrality parameter $\|\xi\|^2$. The likelihood function $L(\lambda)$ for an estimate λ of $\|\xi\|^2$ is given by

$$L(\lambda)=\exp\left(-\frac{\lambda}{2}\right)\sum_{k=0}^{\infty}\frac{(\lambda/2)^k}{k}\frac{w^{(p+2k)/2-1}}{2^{(p+2k)/2}\Gamma((p+2k)/2)}\exp\left(-\frac{w}{2}\right),$$

where $w=\|z\|^2$ and $\lambda\geqq 0$. From the definition of the Bessel function $I_\nu(s)$ (Watson [16], p. 77) this likelihood can be expressed in a compact form

$$L(\lambda)=\frac{1}{2}\exp\left(-\frac{\lambda}{2}\right)(\lambda w)^{-\nu/2}I_\nu((\lambda w)^{1/2})w^\nu\exp\left(-\frac{w}{2}\right),$$

where $\nu=(p-2)/2$. By using the relation $(d/d\sigma)\{\sigma^{-\nu}I_\nu(\sigma)\}=\sigma^{-\nu}I_{\nu+1}(\sigma)$ (Watson [16], p. 79), the first and second order derivatives of the likelihood function with respect to the variable $\sigma=(\lambda w)^{1/2}$ (>0) are obtained as follows:

$$\frac{d}{d\sigma}L(\lambda)=\frac{1}{2}\exp\left(-\frac{\sigma^2}{2w}\right)\sigma^{-\nu}\left\{I_{\nu+1}(\sigma)-\frac{\sigma}{w}I_\nu(\sigma)\right\}w^\nu\exp\left(-\frac{w}{2}\right)$$

(3.4)

$$\frac{d^2}{d\sigma^2}L(\lambda)=\frac{1}{2}\exp\left(-\frac{\sigma^2}{2w}\right)\sigma^{-\nu}\left\{I_{\nu+2}(\sigma)+\left(\frac{1}{\sigma}-2\frac{\sigma}{w}\right)I_{\nu+1}(\sigma)-\left(\frac{1}{w}-\frac{\sigma^2}{w^2}\right)I_\nu(\sigma)\right\}w^\nu\exp\left(-\frac{w}{2}\right).$$

These formulas are useful for the maximum likelihood computation. As will be shown in Section 4, $\sigma I_\nu(\sigma)/I_{\nu+1}(\sigma)$ is strictly increasing with σ and we can see from the above representation of $(d/d\sigma)L(\lambda)$ that the likelihood equation has a unique solution. The square root of the ratio of the maximum likelihood estimate $\hat{\lambda}$ of $\|\xi\|^2$ to $\|z\|^2$ is a function of $w\,(=\|z\|^2)$ only and defines that part of our shrinkage factor required due to the relation between the amplitudes of ξ and z. We will call this quantity the amplitude factor and denote it by

$$\text{AMPFTR}\,(w)=\frac{\hat{\lambda}}{\|z\|}. \tag{3.5}$$

Given $\|z\|^2=w$ the product $\text{AMPFTR}\,(w)\cdot w$ gives the maximum likelihood estimate of $\|\xi\|\cdot\|z\|$ and this is used to define the maximum likelihood estimate of $\text{ANGFTR}\,(\|\xi\|\cdot\|z\|)$ defined by (3.3). The product of the two factors defines the maximum likelihood estimate of ρ_c of (3.1) and is used as the shrinkage factor of our estimator. The new estimator thus obtained will be called the entropy maximizing estimator.

4. Analysis of the new estimator

From (3.4) we have

$$\frac{d}{d\sigma}L(\lambda)=\frac{1}{2}\exp\left(-\frac{\sigma^2+w^2}{2w}\right)\left(\frac{w}{\sigma}\right)^{\nu}\frac{I_{\nu+1}(\sigma)}{w}\left(w-\frac{\sigma I_{\nu}(\sigma)}{I_{\nu+1}(\sigma)}\right),$$

where $\sigma=(w\lambda)^{1/2}$ and $\nu=(p-2)/2$. We have $(d/d\sigma)(\sigma I_{\nu}(\sigma)/I_{\nu+1}(\sigma))=2(\nu+1)\cdot(I_{\nu+1}^2(\sigma)-I_{\nu}(\sigma)I_{\nu+2}(\sigma))/I_{\nu+1}^2(\sigma)$ and as will be shown shortly $I_{\nu+1}^2(s)-I_{\nu}(s)I_{\nu+2}(s)$ is non-negative. This shows that the term $\sigma I_{\nu}(\sigma)/I_{\nu+1}(\sigma)$ is non-decreasing with σ and thus $(d/d\sigma)L(\lambda)$ can change its sign at most only once. Using the asymptotic relation $I_{\nu}(s)=(s/2)^{\nu}/\nu!+o(s^{\nu})$ as s tends to 0, we can show that $(d/d\sigma)L(\lambda)$ can not be positive for $w\leqq 2(\nu+1)$ $(=p)$ and for this case the maximum likelihood estimate of $\|\boldsymbol{\xi}\|^2$ is equal to 0. For the case where w $(=\|z\|^2)$ is greater than p, the likelihood equation takes the form

$$w=\sigma I_{\nu}(\sigma)/I_{\nu+1}(\sigma)\ .$$

Hereafter we adopt the convention to denote the solution of the above equation simply by σ and the maximum likelihood estimate of $\|\boldsymbol{\xi}\|^2$ by λ. Since $\lambda=\sigma^2/w$ we have $w-\lambda=\sigma(I_{\nu}^2-I_{\nu+1}^2)/(I_{\nu}I_{\nu+1})$, where I_{ν} denotes $I_{\nu}(\sigma)$. To show that $w-\lambda\leqq p$ $(=2(\nu+1))$ holds we have only to show that $2(\nu+1)I_{\nu}I_{\nu+1}-\sigma(I_{\nu}^2-I_{\nu+1}^2)$ is non-negative. By using the relations (Watson [16], p. 79)

$$I_{\nu-1}(s)-I_{\nu+1}(s)=\frac{2\nu}{s}I_{\nu}(s) \tag{4.1}$$

and

$$2\frac{d}{ds}I_{\nu}(s)=I_{\nu-1}(s)-I_{\nu+1}(s) \tag{4.2}$$

we can get the equations $2(\nu+1)I_{\nu}I_{\nu+1}-\sigma(I_{\nu}^2-I_{\nu+1}^2)=\sigma(I_{\nu+1}^2-I_{\nu}I_{\nu+2})$ and $(d/ds)(I_{\nu+1}^2(s)-I_{\nu}(s)I_{\nu+2}(s))=2I_{\nu}(s)I_{\nu+2}(s)$. Since $I_{\nu}(0)=0$ for $\nu>0$ and $I_{\nu}(s)\geqq 0$ this completes the desired proof. We can also show that $p-1\leqq w-\lambda$. To show this we use (4.1) and (4.2) and get the equations $\sigma(I_{\nu}^2-I_{\nu+1}^2)-(p-1)I_{\nu}I_{\nu+1}=(\sigma/2)(I_{\nu}^2-I_{\nu-1}I_{\nu+1}-I_{\nu+1}^2+I_{\nu}I_{\nu+2})$ and $(I_{\nu}^2(s)+I_{\nu}(s)I_{\nu+2}(s)-I_{\nu+1}^2(s)-I_{\nu-1}(s)\cdot I_{\nu+1}(s))/(2I_{\nu}^2(s))=(d/ds)(I_{\nu+1}(s)/I_{\nu}(s))$. This last quantity is the derivative of ANGFTR (s) with respect to its argument s. From (3.2) ANGFTR (s) is the mean $\mathrm{E}(t)$ of a random variable t with a probability distribution defined by the density function $C(1-t^2)^{(\nu-1)/2}\exp(st)$ for $|t|\leqq 1$ and 0 elsewhere and (d/ds) ANGFTR (s) is equal to the variance of this distribution and hence non-negative and the desired result follows. We already know that as w tends to p the maximum likelihood estimate λ tends to 0, i.e., $w-\lambda$ tends to p. When w grows indefinitely, using the asymp-

totic relation $I_\nu(s)=\exp(s)(2\pi s)^{-1/2}\{1-(4\nu^2-1)/8s\}+o(1/s)$, it can be shown that $w-\lambda$ tends to $p-1$.

Summarizing the above results we can see that our maximum likelihood estimate λ of $\|\boldsymbol{\xi}\|^2$ lies between $\{w-(p-1)\}^+$ and $\{w-p\}^+$, where $\{\ \}^+$ denotes the positive part, and that it is close to $\{w-p\}^+$ when w is close to or smaller than p and close to $\{w-(p-1)\}^+$ when w is large compared with p^2.

The behavior of ANGFTR can be analyzed analogously. First, the mode of the conditional distribution of $(\boldsymbol{\xi}, z)/(\|\boldsymbol{\xi}\|\cdot\|z\|)$, given $\|z\|$ and $\|\boldsymbol{\xi}\|$, is obtained as

$$\frac{((p-3)^2+4\sigma^2)^{1/2}-(p-3)}{2\sigma},$$

where $\sigma=\|\boldsymbol{\xi}\|\cdot\|z\|$. For the purpose of comparison, we slightly modify this mode and define our approximate angle factors by

$$\text{angftr}(\sigma; i)=\frac{((p-i)^2+4\sigma^2)^{1/2}-(p-i)}{2\sigma} \qquad \text{for } i=0, 1, 2.$$

We also define the corresponding approximate amplitude factors for $i=0, 1, 2$ by

$$\text{ampftr}(w; i)=\left[\left\{1-\frac{p-i}{w}\right\}^+\right]^{1/2}.$$

The product of angftr $(\sigma; i)$ and ampftr $(w; i)$ is equal to $\{1-(p-i)/w\}^+$. For ANGFTR the following inequalities hold:

$$\text{angftr}(\sigma; 1)\geqq\text{ANGFTR}(\sigma)\geqq\text{angftr}(\sigma; 0).$$

For the proof of the first inequality we note the relation ANGFTR (σ) $=I_{\nu+1}(\sigma)/I_\nu(\sigma)=\{\sigma/(2\nu+1)\}\{1-\mathrm{E}(t^2)\}$ where t is the random variable defined in the next to the last paragraph with s replaced by σ. Since $\mathrm{E}(t^2)\geqq(\mathrm{E}(t))^2$ and ANGFTR $(\sigma)=\mathrm{E}(t)$ we have ANGFTR $(\sigma)\leqq\{\sigma/(p-1)\}$ $\cdot\{1-(\text{ANGFTR}(\sigma))^2\}$ and from this inequality follows the first of the above inequalities. For the proof of the second inequality we have only to show that $(\text{ANGFTR}(\sigma))^2+\{(2\nu+2)/\sigma\}\,\text{ANGFTR}(\sigma)\geqq 1$. The left-hand side is equal to $(I_{\nu+1}(\sigma)/I_\nu(\sigma))\{I_{\nu+1}(\sigma)/I_\nu(\sigma)+(2\nu+2)/\sigma\}$. Form the equality $I_{\nu-1}(s)-I_{\nu+1}(s)=(2\nu/s)I_\nu(s)$ we get $\{I_{\nu+1}(\sigma)/I_\nu(\sigma)\}\{I_{\nu+2}(\sigma)/I_{\nu+1}(\sigma)+(2\nu+2)/\sigma\}=1$. We have already shown that $I_{\nu+1}^2(\sigma)\geqq I_\nu(\sigma)I_{\nu+2}(\sigma)$ and thus the desired result follows. It can further be shown that as σ tends to 0 both ANGFTR (σ) and angftr $(\sigma; 0)$ can be approximated by σ/p and as σ tends to infinity both ANGFTR (σ) and angftr $(\sigma; 1)$ can be approximated by $1-(p-1)/(2\sigma)$.

Since our shrinkage factor is obtained as the product of AMPFTR $(\|z\|^2)$ and ANGFTR (σ) defined with $\sigma=\text{AMPFTR}(\|z\|^2)\cdot\|z\|^2$ the results

obtained in this section show that the shrinkage factor always lies between $\{1-p/\|z\|^2\}^+$ and $\{1-(p-1)/\|z\|^2\}^+$ and that it comes closer to the former when $\|z\|^2$ takes small values and to the latter when $\|z\|^2$ takes large values. Our factor is always more shrinking than that of the James-Stein positive part estimator $\{1-(p-2)/\|z\|^2\}^+z$ (Efron and Morris [5]) which is obtained by taking the product of ampftr $(\|z\|^2; 2)$ and angftr $(\sigma; 2)$.

5. Discussions

For the evaluation of the performance characteristics of our new estimator we follow the Bayesian approach adopted by Efron and Morris [5]. We assume the situation where ξ_i's are independently identically distributed as $N(0, A^2)$. For this case the Bayes estimator of $\boldsymbol{\xi}$ is defined by $\boldsymbol{\xi}^0=(1-C)\boldsymbol{z}$ with $C=1/(1+A^2)$. For an estimator $\boldsymbol{\xi}^*$ of $\boldsymbol{\xi}$ the relative savings loss is defined as $\mathrm{RSL}\,(C, \boldsymbol{\xi}^*)=\{\mathrm{E}\,\|\boldsymbol{\xi}^*-\boldsymbol{\xi}\|^2-\mathrm{E}\,\|\boldsymbol{\xi}^0-\boldsymbol{\xi}\|^2\}/\{\mathrm{E}\,\|\boldsymbol{z}-\boldsymbol{\xi}\|^2-\mathrm{E}\,\|\boldsymbol{\xi}^0-\boldsymbol{\xi}\|^2\}$. From Lemma 1 of Efron and Morris [5], for an estimator of the form $\boldsymbol{\xi}^*=(1-\tau(\|\boldsymbol{z}\|^2)/\|\boldsymbol{z}\|^2)\boldsymbol{z}$, we have

$$\mathrm{RSL}\,(C, \xi^*)=\mathrm{E}\left\{\frac{\tau(\|z\|^2)}{C\|z\|^2}-1\right\}^2,$$

where E denotes the expectation with respect to the chi-squared variable $C\|z\|^2$ with the degrees of freedom $p+2$. We can show the following relation for $p\geqq 3$

$$1-\mathrm{RSL}\,(C, \xi^*)=\frac{1}{p(p-2)}\,\mathrm{E}\left\{2\tau\left(\frac{v}{C}\right)v-\tau^2\left(\frac{v}{C}\right)\right\}, \tag{5.1}$$

where E is taken with respect to the distribution of a chi-squared variable v with the degrees of freedom $p-2$. From this result we can see that when $\tau(\|z\|^2)$ is a constant, as is in the original estimator of James and Stein [8], the relative savings loss is independent of C and this allows a uniform comparison of this type of estimators. The situation changes drastically when $\tau(\|z\|^2)$ depends on $\|z\|^2$. If $\tau(\|z\|^2)$ tends to a constant as $\|z\|^2$ grows indefinitely the above result shows that, for C tending to 0, the performance of the estimator is approximated by the estimator obtained by replacing $\tau(\|z\|^2)$ by the constant. When C tends to 1, its upper limit, the performance depends on the details of $\tau(\|z\|^2)$. This is the case where $\boldsymbol{\xi}$ is almost equal to 0 and a severe reduction of z by the shrinkage factor is quite desirable. By the present criterion the shrinkage factor defined by $\tau(\|z\|^2)=p-2$ is better than that defined by $\tau(\|z\|^2)=p-1$, but for their positive part versions, defined by $\tau(\|z\|^2)=p-i$, for $\|z\|^2\geqq p-i$, $\|z\|^2$, otherwise, $(i=1, 2)$, the relation is reversed as C tends to 1. From the analysis of the preceding section we know

that our new estimator is defined with a $\tau(\|z\|^2)$ such that its behavior at $\|z\|^2$ close to or less than p is almost identical to the positive part version of an estimator defined by $\tau(\|z\|^2)=p$. While the behavior of the $\tau(\|z\|^2)$ is close to that of $\tau(\|z\|^2)=p-1$ when $\|z\|^2$ is large. This suggests that $\tau(v/C)$ of our estimator switches to the better choice, at least at the two extreme values of $C=0$ and 1. It is easy to show that for our estimator the right-hand side of (5,1) is positive, i.e., our estimator is better than the maximum likelihood estimator, if $C>1/2$ or $p\geqq 6$. Due to the implicit definition of our shrinkage factor, exact evaluation of the relative savings loss seems rather difficult and extensive Monte Carlo experiments have been performed to check our conjecture that the new estimator will be better than the maximum likelihood estimator, in terms of the relative savings loss, for $p\geqq 3$. The most critical is the case $p=3$, since from the result of James and Stein [8] the estimator defined by $\tau(\|z\|^2)=p$ is already known not to be uniformly better than the ordinary maximum likelihood estimator in this case. One set of numerical results for this case is given in the Table 1. It can be seen that, although the positive part estimator correspond-

Table 1 Comparison of efficiencies for the case $p=3$

A	Average sum of squared errors (sample size=1000)					
	Bayes	MLE	$(p-2)^+$	$(p-1)^+$	$(p)^+$	EME
0.125	0.05	2.99	1.61	0.91	0.53	0.74
0.5	0.59	2.99	1.87	1.31	1.02	1.22
1.0	1.46	2.99	2.27	1.95	1.85	1.98
2.0	2.33	2.99	2.72	2.68	2.77	2.77
5.0	2.84	2.99	2.95	2.96	3.03	2.97
10.0	2.94	2.99	2.98	2.98	3.01	2.99

A: standard deviation of the prior distribution
Bayes: Bayes estimator
MLE: maximum likelihood estimator
$(p-i)^+$: positive part estimator corresponding to $\tau(\|z\|^2)=p-i$
EME: entropy maximizing estimator

ing to $\tau(\|z\|^2)=p$ is obviously showing poor performance compared with the maximum likelihood estimator at large values of A, the sample means of the sum of squared errors of our estimator are, as was expected by the above stated switching behavior, uniformly smaller than those of the ordinary maximum likelihood estimator. In our experiments, for the cases with $p\geqq 4$, our estimator has been almost invariably producing better results than the maximum likelihood estimator. One typical example is shown in Table 2 for the case $p=6$. One by-product of these experiments was the recognition of the practical utility

of our estimator for the case $p=2$. Our estimator performed well at small values of A and the deterioration of the performance at larger values of A seemed quite tolerable for practical applications. The positive part estimator corresponding to $\tau(\|z\|^2)=p-1$ was also performing quite well in this case. A numerical result to show this is given in Table 3. It is interesting to note that in these examples entropy maxi-

Table 2 Comparison of efficiencies for the case $p=6$

A	Average sum of squared errors (sample size=1000)					
	Bayes	MLE	$(p-2)^+$	$(p-1)^+$	$(p)^+$	EME
0.125	0.09	6.10	1.45	1.00	0.69	0.85
0.5	1.20	6.10	2.36	1.99	1.77	1.92
1.0	2.99	6.10	3.83	3.69	3.67	3.76
2.0	4.81	6.10	5.24	5.24	5.32	5.29
5.0	5.82	6.10	5.91	5.91	5.93	5.92
10.0	6.02	6.10	6.04	6.04	6.04	6.04

Table 3 Comparison of efficiencies for the case $p=2$

A	Average sum of squared errors (sample size=1000)					
	Bayes	MLE	$(p-2)^+$	$(p-1)^+$	$(p)^+$	EME
0.125	0.03	2.00	2.00	0.87	0.42	0.67
0.5	0.41	2.00	2.00	1.08	0.74	0.97
1.0	1.01	2.00	2.00	1.51	1.38	1.54
2.0	1.61	2.00	2.00	1.87	1.96	1.94
5.0	1.92	2.00	2.00	2.00	2.12	2.04
10.0	1.98	2.00	2.00	2.01	2.07	2.03

mizing estimator is performing badly compared with the two neighbouring positive part estimators at $A=1.0$. The maximum likelihood estimates of $\|\xi\|^2$ in these examples were obtained by the Newton-Raphson procedure using $\{\|z\|^2-p\}^+$ as the initial values.

Obviously much remains to be done for the clarification of the characteristics of our estimator. Yet it would be quite safe to say that our entropy maximizing approach is producing a meaningful result. The application of the present procedure is not limited to the estimation of the mean of a multivariate normal distribution. Practically it can be applied to any ordinary maximum likelihood estimates of multiple parameters, if only the asymptotic normality of the distribution of the estimates are assured. The only necessary modification is the transformation of the estimates so that they will have a unit variance covariance matrix. It seems that, as an extension of the method of maximum likelihood, the entropy maximizing approach has opened up the possibility

of developing a fairly useful closed system of inference which does not require an ad hoc principle such as Bayes/non-Bayes compromise or unbiasedness.

Acknowledgements

The author is grateful to Dr. K. Tanabe for the very helpful discussions at various stages of the present study. Thanks are also due to Miss E. Arahata for the programming of the Monte Carlo experiments which provided useful information for the analysis reported in the present paper.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In *2nd International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds.), 267-281, Akademiai Kiado, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, AC-**19**, 716-723.

[3] Boltzmann, L. (1877). Uber die Beziehung zwischen dem zweiten Hauptsatz der mechanischen Warmetheorie und der Wahrscheinlichkeitsrechnung respective den Satzen uber das Warmegleichgewicht, *Wiener Berichte*, **76**, 373-435.

[4] Dempster, A. P. (1973). Alternatives to least squares in multiple regression, In *Multivariate Statistical Inference* (D. G. Kabe and R. P. Gupta, eds.), 25-40, North-Holland, Amsterdam.

[5] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach, *J. Amer. Statist. Ass.*, **68**, 117-130.

[6] Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations, *J. Amer. Statist. Ass.*, **70**, 311-319.

[7] Good, I. J. (1965). *The Estimation of Probabilities*, M.I.T. Press, Cambridge.

[8] James, W. and Stein, C. M. (1961). Estimation with quadratic loss function, *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, **1**, 361-379.

[9] Kullback, S. (1959). *Information and Statistics*, Wiley, New York.

[10] Lindley, D. V. (1962). Discussion of a paper by C. Stein, *J. R. Statist. Soc.*, B, **24**, 285-296.

[11] Sanov, I. N. (1961). On the probability of large deviations of random variables, *IMS and AMS Selected Translation in Mathematical Statistics and Probability*, 1, 213-244.

[12] Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, **1**, 197-206.

[13] Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution, *J. R. Statist. Soc.*, B, **24**, 265-285.

[14] Stein, C. M. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs, In *Festschrift for J. Neyman*: *Research Papers in Statistics* (F. N. David, ed.), 351-366, Wiley, New York.

[15] Vincze, I. (1974). On the maximum probability principle in statistical physics, In *Progress in Statistics*, **2**, (J. Gani, K. Sarkadi and I. Vince, eds.), 869-893, North-Holland, Amsterdam.

[16] Watson, G. N. (1922). *A Treatise on the Theory of Bessel Functions*, Cambridge University Press.